Rough Set Theory as a Data Mining Technique: A Case Study in Epidemiology and Cancer Incidence Prediction

Chelly Dagdia, Zaineb; Zarges, Christine; Schannes, Benjamin; Micalef, Martin; Galiana, Lino; Rolland, Benoît; de Fresnoye, Olivier; Benchoufi, Mehdi

Published in:
Machine Learning and Knowledge Discovery in Databases

DOI:
10.1007/978-3-030-10997-4_27

Publication date:
2019

Citation for published version (APA):
Chelly Dagdia, Z., Zarges, C., Schannes, B., Micalef, M., Galiana, L., Rolland, B., de Fresnoye, O., & Benchoufi, M. (2019). Rough Set Theory as a Data Mining Technique: A Case Study in Epidemiology and Cancer Incidence Prediction. In U. Brefeld, E. Curry, E. Daly, B. MacNamee, A. Marascu, F. Pinelli, M. Berlingerio, & N. Hurley (Eds.), *Machine Learning and Knowledge Discovery in Databases: ECML PKDD 20-18* (pp. 440-455). (Lecture Notes in Computer Science ; Vol. 11053). Springer Nature. https://doi.org/10.1007/978-3-030-10997-4_27

Rough Set Theory as a Data Mining Technique: A Case Study in Epidemiology and Cancer Incidence Prediction*

Zaineb Chelly Dagdia[✉][1,2], Christine Zarges[1], Benjamin Schannes[3], Martin Micalef[4], Lino Galiana[5], Benoît Rolland[6], Olivier de Fresnoye[7], and Mehdi Benchoufi[8]

[1] Department of Computer Science, Aberystwyth University, Aberystwyth, United Kingdom
{zaineb.chelly,c.zarges}@aber.ac.uk
[2] LARODEC, Institut Supérieur de Gestion de Tunis, Tunis, Tunisia
chelly.zaineb@gmail.com
[3] Department of Statistics, ENSAE, 5 avenue Henry Le Chatelier, 91120 Palaiseau, France, benjamin.schannes@gmail.com
[4] Actuaris, 13/15 boulevard de la Madeleine, 75001 Paris, France
martin.micalef@gmail.com
[5] ENS Lyon, 15 parvis René Descartes, 69342 Lyon Cedex 07, France
lino.galiana@ens-lyon.fr
[6] Altran Technologies S.A., 96 rue Charles de Gaulle, 92200 Neuilly-sur-Seine, France
benoit.rolland@free.fr
[7] Coordinateur Scientifique Programme Épidemium, France, olivier@epidemium.cc
[8] Centre d'Épidémiologie Clinique, Hôpital Hôtel Dieu, Assistance Publique-Hôpitaux de Paris, France - Faculté de Médecine, Université Paris Descartes and INSERM UMR1153, Paris, France - Coordinateur Scientifique Programme Épidemium, mehdi.benchoufi@aphp.fr

Abstract. A big challenge in epidemiology is to perform data preprocessing, specifically feature selection, on large scale data sets with a high dimensional feature set. In this paper, this challenge is tackled by using a recently established distributed and scalable version of Rough Set Theory (RST). It considers epidemiological data that has been collected from three international institutions for the purpose of cancer incidence prediction. The concrete data set used aggregates about 5 495 risk factors (features), spanning 32 years and 38 countries. Detailed experiments demonstrate that RST is relevant to real world big data applications as

* This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 702527. This work was based on a first version of a database provided by the OpenCancer organization, part of Épidemium—a data challenge oriented and community—based open science program. Additional thanks go to the Épidemium group, Roche, La Paillasse and to the Supercomputing Wales project, which is part-funded by the European Regional Development Fund via the Welsh Government.

it can offer insights into the selected risk factors, speed up the learning process, ensure the performance of the cancer incidence prediction model without huge information loss, and simplify the learned model for epidemiologists.

Keywords: Big Data · Rough Set Theory · Feature Selection · Epidemiology · Cancer Incidence Prediction · Application

1 Introduction

Epidemiology is a sub-field of public health that looks to determine where and how often disease occur and why. It is more formally defined as the study of distributions (patterns) and determinants (causes) of health related states or events within a specified human population, and the application of this study to managing health problems [4]. The ultimate goal of epidemiology is to apply this knowledge to the control of disease through prevention and treatment, resulting in the preservation of public health.

In this context, epidemiologists study chronic diseases such as arthritis, cardiovascular disease such as heart attacks and stroke, cancer such as breast and colon cancer, diabetes, epilepsy and obesity problems. To conduct such studies, one of the most important considerations is the source and content of data, as this will often determine the quality of the results. As a general rule, the larger the data, the more accurate the results, since a larger sample is less likely to, by chance, generate an estimate different from the truth in the full population. This leads epidemiologists to deal with large amounts of data, big data, which is however not a feasible task for them [3]. Hence, to assist epidemiologists in dealing with such large amounts of data, data analysis has become one of the major research focuses in epidemiology and specifically for the epidemiology of cancer, colon cancer, which is our main focus. More precisely, data analysis assists epidemiologists to investigate and describe the determinants and distribution of disease, disability, and other health outcomes and develop the means for prevention and control. From a technical perspective, data analysis generally comprises a number of processes that may include data collection, data (pre)-processing and feature reduction, data cleansing, and data transformation and modeling with the goal of discovering useful information, suggesting conclusions, and supporting decision making; all these tasks can be achieved via the use of adequate machine learning techniques.

Meanwhile, in epidemiology, feature reduction is a main point of interest across the various steps of data analysis and focusing on this phase is crucial as it often presents a source of potential information loss. Many techniques were proposed in the literature [2] to achieve the task of feature reduction and they can be categorized into two main categories: techniques that transform the original meaning of the features, called "transformation-based approaches" or "feature extraction approaches", and semantic-preserving techniques that attempt to retain the meaning of the original feature set, known as "selection-based approaches" [11]. Within the latter category a further partitioning can be

defined where the techniques are classified into filter approaches and wrapper approaches. The main difference between the two branches is that wrapper approaches include a learning algorithm in the feature subset evaluation, and hence they are tied to a particular induction algorithm. In this work, we mainly focus on the use of a feature selection technique, specifically a filter technique, instead of a feature extraction technique. This is crucial to preserve the semantics of the features in the context of cancer incidence prediction as results should be interpretable and understandable by epidemiologists.

Yet, the adaptation of feature selection techniques for big data problems may require the redesign of these algorithms and their inclusion in parallel and distributed environments. Among the possible alternatives is the MapReduce paradigm [13] introduced by Google which offers a robust and efficient framework to address the analysis of big data. Several recent works have focused on the parallelization of machine learning tools using the MapReduce approach [12,15,14,16]. Recently, new and more flexible workflows have appeared to extend the standard MapReduce approach, mainly Apache Spark [18], which has been successfully applied over various data mining and machine learning problems [18]. With the aim of choosing the most relevant and pertinent subset of features, a variety of feature selection techniques were proposed to deal with big data in a distributed way [20]. Nevertheless, most of these techniques suffer from some shortcomings. For instance, they usually require expert knowledge for the task of algorithm parameterization or noise levels to be specified beforehand and some simply rank features leaving the user to choose their own subset. There are some techniques that need the user to specify how many features are to be chosen, or they must supply a threshold that determines when the algorithm should terminate. All of these require the expert or the user to make a decision based on their own (possibly faulty) judgment. To overcome the limitations of the state-of-the-art methods, it is interesting to look for a filter method that does not require any external or supplementary information to function properly. Rough Set Theory (RST) can be used as such a technique [6]. RST, as a powerful feature selection technique, has made many achievements in many applications such as in decision support, engineering, environment, banking, medicine and others [19]. In this study, we focus on the use of RST as a data mining technique within a case study in epidemiology and cancer incidence prediction.

The rest of this paper is structured as follows. Section 2 reviews the fundamentals of epidemiology. Section 3 introduces the basic concepts of rough set theory for feature selection. Section 4 details the application in epidemiology and cancer incidence prediction via the use of a distributed algorithm based on rough sets for large-scale data pre-processing. The experimental setup is introduced in Section 5. The results of the performance analysis are discussed in Section 6 and conclusions are presented in Section 7.

2 Epidemiology: Concepts and Context Design

2.1 Distribution and Determinants

Epidemiology is concerned with the study of the distribution of a disease based on a set of "frequency" and "pattern" of health events in a population. The frequency refers on one hand to the number of health events such as the number of cases of diabetes or cancer in a population, and on the other hand to the link of that number to the size of the human population. The resulting ratio permits epidemiologists to compare disease occurrence across diverse populations. Pattern denotes the occurrence of health-related events by person, place, and time. Personal patterns comprise demographic factors that may be tied to risk of sickness, injury, or disability such as age, sex, marital status, social class, racial group, occupation, as well as behaviors and environmental exposures. Place patterns include geographic disparity, urban/suburban/rural variances, and location of work sites or schools. Time patterns can be annual, seasonal, weekly, daily, hourly, or any other breakdown of time that may effect disease or injury occurrence [10]. Moreover, epidemiology is concerned with the search for determinants. These are the factors that precipitate disease. Formally, determinants can be defined as any factor, whether event, characteristic, or other definable entity, that brings about a change in a health condition or other defined characteristic [9]. Epidemiologists assume that a disease does not arise haphazardly in a population, but it occurs when a set of accumulation of risk factors or determinants subsists in an individual. To look for these determinants, epidemiologists use epidemiological studies to understand and answer the "Why" and "How" of such events. For instance, they assess whether groups with dissimilar rates of disease diverge in their demographic characteristics, genetic or immunologic make-up, or any other so-called potential risk factors. Ideally, the findings provide sufficient evidence to direct prompt and effective public health control and prevention measures [10].

2.2 Population and Samples

An epidemiological study involves the collection, analysis and interpretation of data from a human population. The population about which epidemiologists wish to draw conclusions is called the "target population". In many cases, this is defined according to geographical criteria or some political boundaries. The specific population from which data are collected is called the "study population". It is a question of judgment whether results of the study population may be used to draw accurate conclusions about the target population. Most of the epidemiological studies use study populations that are based on geographical, institutional or occupational definitions. Another way of classifying the study of population is by the stage of the disease, i.e., a population that is diseased, disease-free or a mixture [4]. On the other hand, a sample is any part of the fully defined population. A syringe full of blood drawn from the vein of a patient is a sample of all the blood in the patient's circulation at the moment. Similarly, 100

patients suffering from colon cancer is a sample of the population of all the patients suffering from colon cancer. To make accurate inferences, the sample has to be properly chosen, representative, and the inclusion and exclusion criteria should be well defined as well. A representative sample is one in which each and every member of the population has an equal and mutually exclusive chance of being selected [8].

2.3 Incidence and Prevalence

Epidemiology often focuses on measuring the occurrence of disease in populations. The basic measures of disease frequency in epidemiology are "incidence" and "prevalence". Incidence is the number of new cases of disease in a population occurring over a defined period of time. Another important measure of disease incidence is incidence rate, which gauges how fast disease occurs in the population by measuring the number of new cases emerging as a function of time. Prevalence, on the other hand, measures the number of existing cases, both new cases and cases that have been diagnosed in the past, in a population at any given point in time. By using these measures, epidemiologists can determine the frequency of disease within populations, and compare differences in disease risk among populations [4].

3 Rough Set Theory

Rough Set Theory (RST) [17,1] is considered to be a formal approximation of the conventional set theory, which supports approximations in decision making. It provides a filter-based technique by which knowledge may be extracted from a domain in a concise way, retaining the information content whilst reducing the amount of knowledge involved [6]. This section focuses mainly on highlighting the fundamentals of rough set theory for feature selection.

3.1 Preliminaries of Rough Set Theory

In rough set theory, an *information table* is defined as a tuple $T = (U, A)$ where U and A are two finite, non-empty sets with U the *universe* of primitive objects and A the set of attributes. Each attribute or feature $a \in A$ is associated with a set V_a of its value, called the *domain* of a. We may partition the attribute set A into two subsets C and D, called *condition* and *decision* attributes, respectively.

Let $P \subset A$ be a subset of attributes. The indiscernibility relation, denoted by $IND(P)$, is the central concept of RST and it is an equivalence relation, which is defined as: $IND(P) = \{(x, y) \in U \times U : \forall a \in P, a(x) = a(y)\}$, where $a(x)$ denotes the value of feature a of object x. If $(x, y) \in IND(P)$, x and y are said to be *indiscernible* with respect to P. The family of all equivalence classes of $IND(P)$, referring to a partition of U determined by P, is denoted by $U/IND(P)$. Each element in $U/IND(P)$ is a set of indiscernible objects with respect to P. The

equivalence classes $U/IND(C)$ and $U/IND(D)$ are called *condition* and *decision* classes, respectively. For any concept $X \subseteq U$ and attribute subset $R \subseteq A$, X could be approximated by the R-*lower* approximation and R-*upper* approximation using the knowledge of R. The lower approximation of X is the set of objects of U that are surely in X, defined as: $\underline{R}(X) = \bigcup\{E \in U/IND(R) : E \subseteq X\}$. The upper approximation of X is the set of objects of U that are possibly in X, defined as: $\overline{R}(X) = \bigcup\{E \in U/IND(R) : E \cap X \neq \emptyset\}$. The concept defining the set of objects that can possibly, but not certainly, be classified in a specific way is called the *boundary region*, which is defined as: $BND_R(X) = \overline{R}(X) - \underline{R}(X)$. If the boundary region is empty, that is $\overline{R}(X) = \underline{R}(X)$, concept X is said to be R-*definable*; otherwise X is a *rough set* with respect to R. The *positive region* of decision classes $U/IND(D)$ with respect to condition attributes C is denoted by $POS_c(D)$ where $POS_c(D) = \bigcup \overline{R}(X)$. The positive region $POS_c(D)$ is a set of objects of U that can be classified with certainty to classes $U/IND(D)$ employing attributes of C. In other words, the positive region $POS_c(D)$ indicates the union of all the equivalence classes defined by $IND(P)$ that each for sure can induce the decision class D. Based on the positive region, the *dependency of attributes* measuring the degree k of the dependency of an attribute c_i on a set of attributes C is defined as: $k = \gamma(C, c_i) = |POS_C(c_i)|/|U|$. Based on these basics, RST defines two important concepts for feature selection, which are the $Core$ and the $Reduct$.

3.2 Reduction Process

RST aims at choosing the smallest subset of the conditional feature set so that the resulting reduced data set remains consistent with respect to the decision feature. To do so, RST defines the Reduct and the Core concepts. In rough set theory, a subset $R \subseteq C$ is said to be a D-*reduct* of C if $\gamma(C, R) = \gamma(C)$ and there is no $R' \subset R$ such that $\gamma(C, R^{'}) = \gamma(C, R)$. In other words, the $Reduct$ is the minimal set of selected attributes preserving the same dependency degree as the whole set of attributes. Meanwhile, rough set theory may generate a set of reducts, $RED_D^F(C)$, from the given information table. In this case, any reduct from $RED_D^F(C)$ can be chosen to replace the initial information table. The second concept, the $Core$, is the set of attributes that are contained by all reducts, defined as $CORE_D(C) = \bigcap RED_D(C)$ where $RED_D(C)$ is the D-reduct of C. Specifically, the $Core$ is the set of attributes that cannot be removed from the information system without causing collapse of the equivalence-class structure. This means that all attributes present in the $Core$ are indispensable.

4 Application

4.1 Data Sources

The OpenCancer[9] organization gathers people working on cancer prediction issues. Their aim is to provide tools aimed at helping health authorities to take

[9] `https://github.com/orgs/EpidemiumOpenCancer/`

public policy decisions in terms of cancer prevention. OpenCancer has linked and merged data from the World Health Organization (WHO)[10], World Bank (WB)[11], the International Labour Organization (ILO)[12] and the Food and Agriculture Organization (FAO)[13] of the United Nations to build a large data set covering 38 countries and many regions within these countries between 1970 and 2002. For this application, OpenCancer provided a first version of the database restricted to the WHO, WB and FAO sources. Each row is characterized by a 5-tuple (cancer type, country, gender, ethnicity, age group) and 5 495 features. For this application the single cancer type, which has been considered, is the colon cancer.

4.2 Data Pre-processing

Data Cleaning The first version of this sub-database suffers from a vast number of missing cells due to the lack of information in the available repositories. To fix this issue prior to running any learning model, OpenCancer had discarded every feature exhibiting a missing data ratio higher than 50 % and imputed other missing data with a standard mean strategy. The resulting database—merged from both FAO and WB, including the incidence provided from WHO—includes 3 365 risk factors (features) and 45 888 records. Each record, seen as a population, is identified via a 6-tuple defined as {Sex, Age group, Country, Region, Ethnicity, Year}. To measure the occurrence of the colon cancer disease in the population, the number of new cases of the disease within a population occurring over 1970 and 2002 is used, referring to the incidence measure.

Feature Selection Once the consistent database is ready for use, a feature selection step is performed. To deal with the large amount of the epidemiological data, a distributed version of rough set theory for feature selection [5], named Sp-RST, is used. Sp-RST is based on a parallel programming design that allows to tackle big data sets over a cluster of machines independently from the underlying hardware and/or software. To select the most important risk factors from the input consistent database, and for the purpose of colon cancer incidence prediction, Sp-RST proceeds as follows:

Problem formalization Technically, the epidemiological database is first stored in an associated Distributed File System (DFS) that is accessible from any computer of the used cluster. To operate on the given DFS in a parallel way, a Resilient Distributed Data set (RDD) is created. We may formalize the latter as a given information table defined as T_{RDD}, where the universe $U = \{x_1, \ldots, x_N\}$ is the set of data items reflecting the population and is identified as a 6-tuple defined as {Sex, Age group, Country, Region, Ethnicity, Year}. The conditional

[10] http://www.who.int/en/
[11] http://www.worldbank.org/
[12] http://www.ilo.org/global/lang–en/index.htm
[13] http://www.fao.org/home/fr/

attribute set $C = \{c_1, \ldots, c_V\}$ contains every single feature of the T_{RDD} information table, and presents the risk factors. The decision attribute D of our learning problem corresponds to the class (label) of each T_{RDD} sample. It has continuous values d and refers to the incidence of the colon cancer. The condition attribute feature D is defined as follows: $D = \{\text{Typology}_1, \ldots, \text{Typology}_I\}$. The conditional attribute set C presents the pool from where the most convenient risk factors will be selected.

Feature selection process For feature selection, the given T_{RDD} information table is partitioned first into m data blocks based on splits from the conditional attribute set C. Hence, $T_{RDD} = \bigcup_{i=1}^{m}(C_r)T_{RDD_{(i)}}$, where $r \in \{1, \ldots, V\}$. Each $T_{RDD_{(i)}}$ is constructed based on r random features selected from C, where $\forall T_{RDD_{(i)}} : \nexists\{c_r\} = \bigcap_{i=1}^{m} T_{RDD_{(i)}}$.

Within a distributed implementation, Sp-RST is applied to every single $T_{RDD_{(i)}}$ so that at the end all the intermediate results will be gathered from the different m partitions. Specifically, Sp-RST will first compute the indiscernibility relation for the decision class defined as $IND(D) : IND(d_i)$. More precisely, Sp-RST will calculate the indiscernibility relation for every decision class d_i by gathering the same T_{RDD} data items, which are defined in the universe $U = \{x_1, \ldots, x_N\}$ and which belong to the same class d_i. This task is independent from the m generated partitions and, as the result, depends on the data items class and not on the features. Once achieved, the algorithm generates the m random $T_{RDD_{(i)}}$ as previously explained. Then, and within a specific partition, Sp-RST creates all the possible combinations of the C_r set of features, computes the indiscernibility relation for every generated combination $IND(AllComb_{(C_r)})$ and calculates the dependency degrees $\gamma(C_r, AllComb_{(C_r)})$ of each feature combination. Then, Sp-RST looks for the maximum dependency value among all $\gamma(C_r, AllComb_{(C_r)})$. The maximum dependency reflects on one hand the dependency of the whole feature set (C_r) representing the $T_{RDD_{(i)}}$ and on the other hand the dependency of all the possible feature combinations satisfying the constraint $\gamma(C_r, AllComb_{(C_r)}) = \gamma(C_r)$. The maximum dependency is the baseline value for feature selection. Then, Sp-RST keeps the set of all combinations having the same dependency degrees as the selected baseline. In fact, at this stage Sp-RST removes in each computation level the unnecessary features that may affect negatively the performance of any learning algorithm.

Finally, Sp-RST keeps the set of combinations having the minimum number of features by satisfying the full reduct constraints discussed in Section 3: $\gamma(C_r, AllComb_{(C_r)}) = \gamma(C_r)$ while there is no $AllComb'_{(C_r)} \subset AllComb_{(C_r)}$ such that $\gamma(C_r, AllComb'_{(C_r)}) = \gamma(C_r, AllComb_{(C_r)})$. Each combination satisfying this condition is considered as a viable minimum reduct set. The attributes of the reduct set describe all concepts in the original training data set $T_{RDD_{(i)}}$.

The output of each partition is either a single reduct $RED_{i_{(D)}}(C_r)$ or a family of reducts $RED^{F}_{i_{(D)}}(C_r)$. Based on the RST preliminaries previously mentioned in Section 3, any reduct of $RED^{F}_{i_{(D)}}(C_r)$ can be used to represent the $T_{RDD_{(i)}}$ information table. Consequently, if Sp-RST generates only

one reduct, for a specific $T_{RDD_{(i)}}$ block, then the output of this feature selection phase is the set of the $RED_{i_{(D)}}(C_r)$ features. These features reflect the most informative ones among the C_r attributes resulting a new reduced $T_{RDD_{(i)}}$, $T_{RDD_{(i)}}(RED)$, which preserves nearly the same data quality as its corresponding $T_{RDD_{(i)}}(C_r)$ that is based on the whole feature set C_r. On the other hand, if Sp-RST generates a family of reducts then the algorithm randomly selects one reduct among $RED^F_{i_{(D)}}(C_r)$ to represent the corresponding $T_{RDD_{(i)}}$. This random choice is justified by the same priority of all the reducts in $RED^F_{i_{(D)}}(C_r)$. In other words, any reduct included in $RED^F_{i_{(D)}}(C_r)$ can be used to replace the $T_{RDD_{(i)}}(C_r)$ features. At this stage, each i data block has its output $RED_{i_{(D)}}(C_r)$ referring to the selected features. However, since each $T_{RDD_{(i)}}$ is based on distinct features and with respect to $T_{RDD} = \bigcup_{i=1}^{m}(C_r)T_{RDD_{(i)}}$ a union of the selected feature sets is required to represent the initial T_{RDD}; defined as $Reduct_m = \bigcup_{i=1}^{m} RED_{i_{(D)}}(C_r)$. In order to ensure the performance of Sp-RST while avoiding considerable information loss, the algorithm runs over N iterations on the T_{RDD} m data blocks and thus generates N $Reduct_m$. Hence, at the end an intersection of all the obtained $Reduct_m$ is needed; defined as $Reduct = \bigcap_{n=1}^{N} Reduct_m$.

By removing irrelevant and redundant features, Sp-RST can reduce the dimensionality of the data from $T_{RDD}(C)$ to $T_{RDD}(Reduct)$. More precisely, Sp-RST was able to reduce the considered epidemiological database from 3 364 risk factors to only around 840 features. The pseudo-code of Sp-RST as well as details related to each of its distributed tasks can be found in [5].

4.3 Predictive Modeling

Accurately evaluating colon cancer risk in average and high-risk populations or individuals and determining colon cancer prognosis in patients are essential for controlling the suffering and death due to colon cancer. From a general perspective, cancer prediction models offer a significant approach to assessing risk and prognosis by detecting populations and individuals at high-risk, easing the design and planning of clinical cancer trials, fostering the development of benefit-risk indices, and supporting estimates of the population burden and cost of cancer. Models also may aid in the evaluation of treatments and interventions, and help epidemiologists make decisions about treatment and long-term follow-up care [7]. In this concern, for colon cancer incidence prediction, the distributed version of the Random Forest Regression model[14] is used.

5 Experimental Setup

5.1 Experimental plan, testbed and tools

Our experiments are performed on the High Performance Computing Wales platform (HPC Wales), which provides a distributed computing facility. Under

[14] org.apache.spark.ml.regression.{RandomForestRegressionModel, RandomForestRegressor}

this testbed, we used 12 dual-core Intel Westmere Xeon X5650 2.67 GHz CPUs and 36GB of memory to test the performance of Sp-RST, which is implemented in Scala 2.11 within Spark 2.1.1. The main aim of our experimentation is to demonstrate that RST is relevant to real world big data applications as it can offer insights into the selected risk factors, speed up the learning process, ensure the performance of the colon cancer incidence prediction model without huge information loss, and simplify the learned model for epidemiologists.

5.2 Parameters settings

As previously mentioned, we use the Random Forest Regression implementation provided in the Spark framework with the following parameters: maxDepth=5, numTrees=20, featureSubsetStrategy='all' and impurity='variance'. The algorithm automatically identifies categorical features and indexes them. The database is split into training and test sets where 30% of the database is held out for testing. Meanwhile, for the Sp-RST settings, we set the number of partitions to 841 partitions; generating 4 features per partition (based on preliminary experiments). We run the settings on 8 nodes on HPC Wales. For the purpose of this study we set the number of iterations of Sp-RST to 10.

6 Results and Discussion

6.1 Categories of Selected Risk Factors

Recall that Sp-RST runs over 10 iterations and that at its last algorithmic stage an intersection of the generated reducts at each iteration is made. However, for the considered data set, this intersection is empty. Hence, we have modified the algorithm to return all 10 different reducts to be presented to the epidemiologists. In the following, we present two different types of results: averages accumulated over the 10 reduced datasets and separate numbers for each of the iterations performed.

Both parts of the data set (FAO, World Bank) contain 9 different categories of risk factors as shown in Table 1. Here, we list the number of different factors in each category, the average number of factors selected over the 10 iterations of Sp-RST and the corresponding percentages. From Table 1, and based on the WB database, we notice that there is only a small variation in the distribution of the selected risk factors. Exceptions are the gender and poverty risk factors, which are not selected by the algorithm. The same comments can be made for the FAO database where the food security risk factor does not appear in any of the selected feature sets. Epidemiologists confirm that these results are quite expected. This demonstrates that our method is able to select the most interesting features to keep—the key risk factors.

We depict the ratios of each category within the set of selected risk factors for each database (FAO & WB) in Figure 1, and for an overall view, the distribution of the categories of the combined data sets is presented in Figure 2.

	#Risk Factors	#selected (average)	% selected
WB: Education	31	7.4	23.87%
WB: Environment	78	19.7	25.26%
WB: Health	62	14.3	23.06%
WB: Infrastructure	19	4.9	25.79%
WB: Economy	141	34.1	24.18%
WB: Public Sector	41	8.5	20.73%
WB: Gender	0	0	-
WB: Social Protection & Labor	40	9.6	24.00%
WB: Poverty	0	0	-
WB: TOTAL	412	98.5	23.91%
FAO: Production	378	90.7	23.99%
FAO: Emissions	803	201.1	25.04%
FAO: Employment	6	1.4	23.33%
FAO: Environment	17	4.4	25.88%
FAO: Commodity	938	234.8	25.03%
FAO: Inputs	153	36	23.53%
FAO: Food Balance	70	20	28.57%
FAO: Food Supply	587	153.7	26.18%
FAO: Food Security	0	0	-
FAO: TOTAL	2952	742.1	25.14%
TOTAL	3364	840.6	24.99%

Table 1. Overview of the data set and the selected risk factors.

Based on these figures, epidemiologists confirmed again that the selected risk factors are expected to appear in each of their corresponding databases (though potentially with a different overall distribution). This again supports that Sp-RST can determine the key risk factors among a large set of features. Meanwhile, epidemiologists highlighted that a higher average or proportion does not necessarily mean that a risk factor is more important than another. Indeed, no firm conclusions on the influence of one factor on the colon incidence prediction can be drawn based on this information, only. Thus, from an epidemiological perspective, the risk factors selected by Sp-RST should be further coupled with other sources of data to complete the analysis and to be able to draw specific conclusions.

We now investigate the selected risk factors per iteration (for all the 10 Sp-RST iterations). Each iteration of SP-RST reflects a possible reduced set of risk factors on which the prediction of the colon cancer incidence can be made. The categories of the selected risk factors in the FAO data set, in the WB database and for the combined database, separately for each of the 10 iterations are presented in Figure 3, Figure 4 and Figure 5, respectively.

Based on Figure 3, Figure 4 and Figure 5, we can see that the risk factors partially overlap within the 10 iterations. This might be interesting from an epidemiological point of view as it can influence the consideration of other pos-

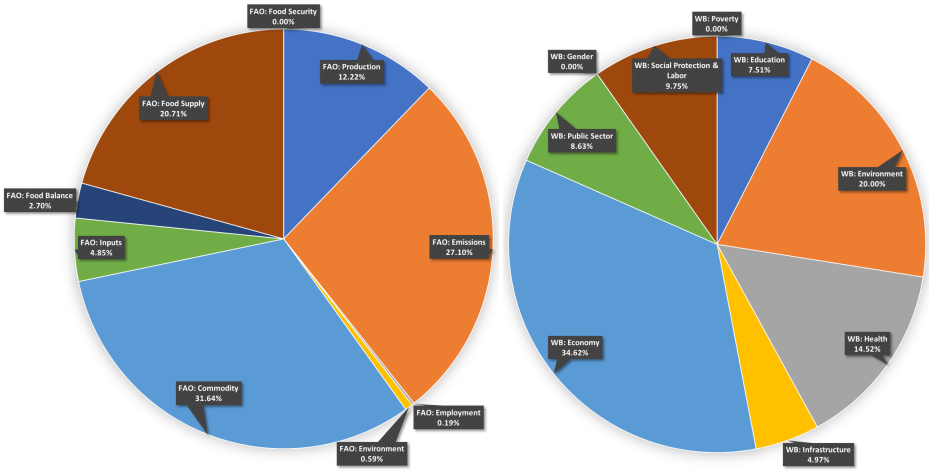

Fig. 1. Distribution of the categories of risk factors selected by our proposed method; split by data set: FAO (left) and World Bank (right).

sible risk factors, which appear with different distributions. Indeed, the overlap between the selected risk factors from one iteration to another may call the attention of the epidemiologist in cases where a firm decision is taken with respect to a specific risk factor. These results are considered to be very important for the epidemiologists as they help them in the decision making process.

6.2 Evaluation of Regression

We use four different metrics to compare the obtained random forest regression models for the original data set and the reduced data sets produced by Sp-RST. Let p_i denote the predicted value of the i-th data item in the test data and v_i its actual value. We call the difference $e_i = v_i - p_i$ the sample error. We consider:

- Mean Absolute Error: $\sum_{i=0}^{n} |e_i|/n$
- Mean Squared Error: $\sum_{i=0}^{n} (e_i)^2|/n$
- Root Mean Squared Error, the square root of the mean squared error
- Coefficient of Determination (R^2): $1 - \sum_{i=0}^{n}(e_i)^2 / \sum_{i=0}^{n}(v_i - \bar{v}_i)^2$, where $\bar{v}_i$ denotes the average of the v_i

For the first three metrics, smaller values indicate a better model. For the R2 metric values between 0 and 1 are obtained, where 1 indicates a perfect model and 0 indicates a trivial model that always predicts the average of the training samples.

Our results for all four metrics are summarized in Table 2. We see that the results are very similar, but slightly better for the original data set. Wilcoxon rank sum tests did not reveal any statistical significance at standard confidence level 0.05 as indicated by the p-values in Table 2. We conclude that the quality

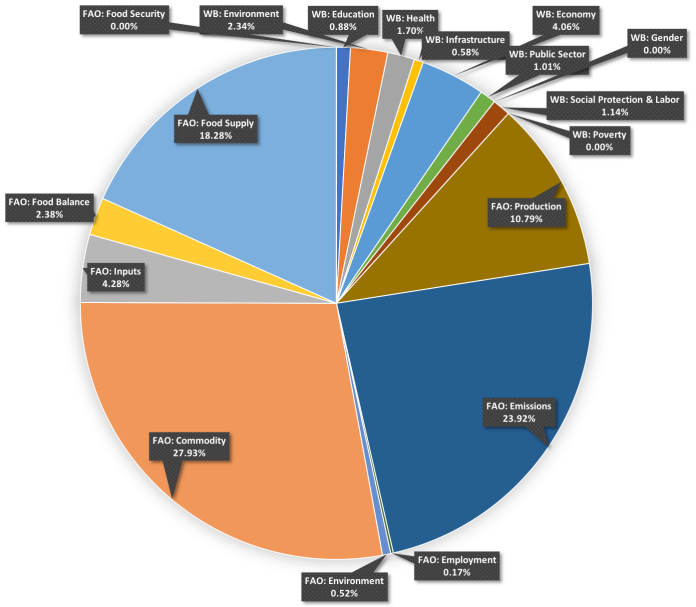

Fig. 2. Distribution of the categories of the combined data set.

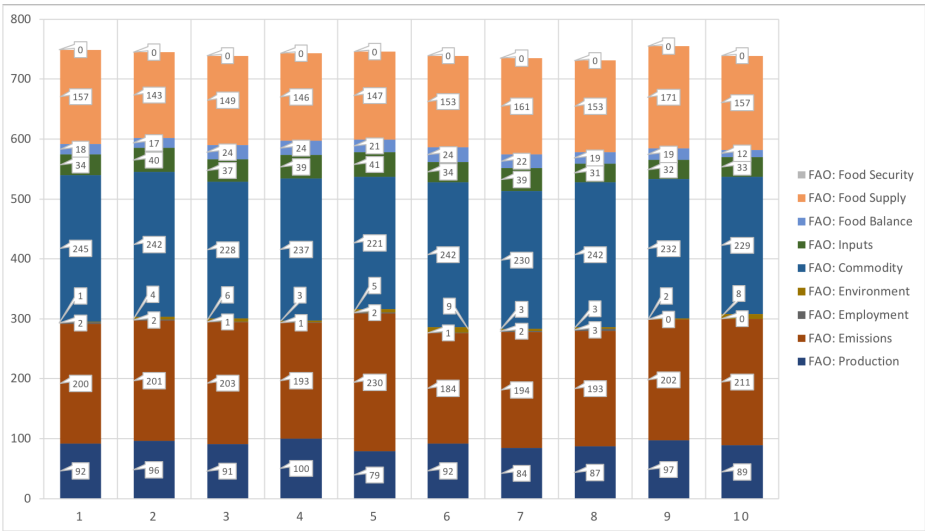

Fig. 3. Categories of the selected risk factors in the FAO data set for each of the 10 iterations.

of the obtained regression models is comparable. However, the reduced data set improves the execution time to determine the regression model considerably (by

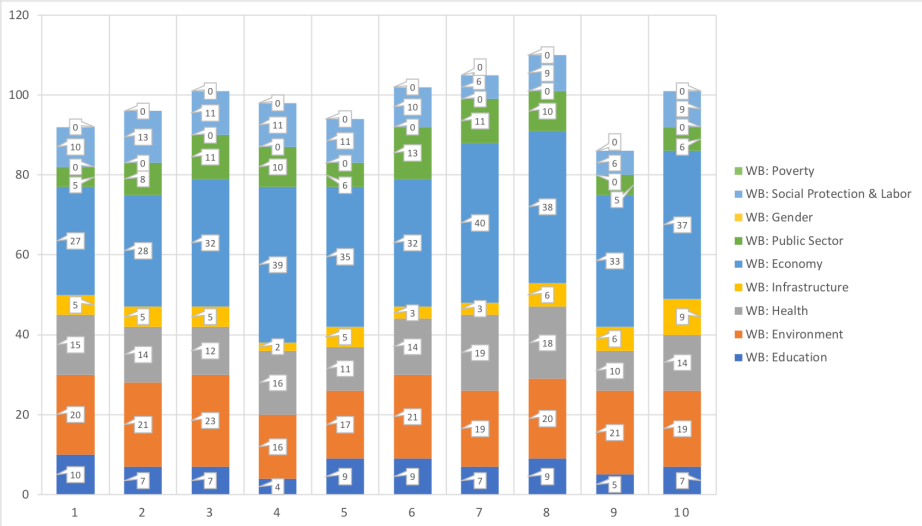

Fig. 4. Categories of the selected risk factors in the World Bank data set for each of the 10 iterations.

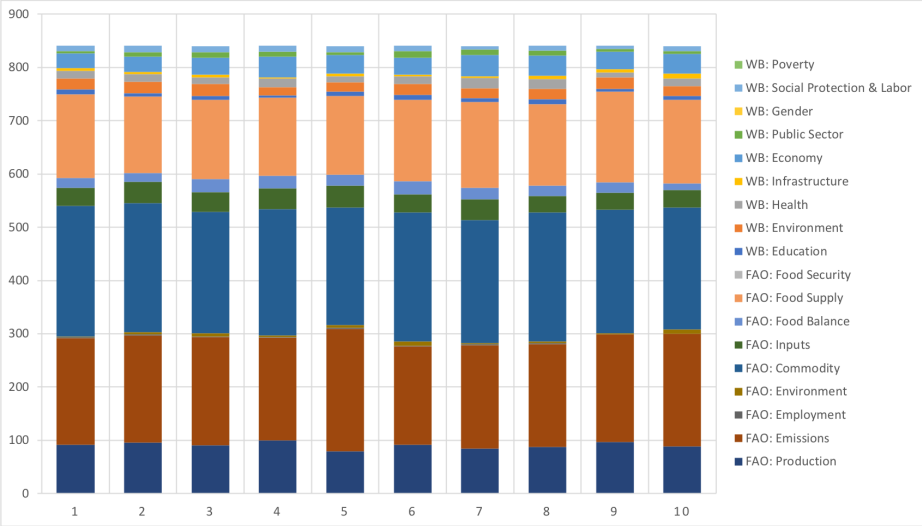

Fig. 5. Categories of the selected risk factors in the combined data set for each of the 10 iterations.

almost a factor of 5). Moreover, a data set with only around 840 risk factors is much easier to interpret and handle by epidemiologists as discussed in the previous section. We therefore argue that the reduction process is appropriate in the considered context.

	mean (Sp-RST)	std (Sp-RST)	mean (original)	sd (original)	p-value
MAE	12.51929	0.167719	12.38475	0.174458	0.1488
MSE	501.4551	20.19929	481.8668	18.7318	0.08688
RMSE	22.38885	0.4488141	21.94872	0.4250894	0.08688
R2	0.2531238	0.02098402	0.2718285	0.01256834	0.1884
Time (s)	78.84557	10.81764	381.9735	5.647821	0.0003666

Table 2. Evaluation of the random forest regression model using Root Mean Squared Error (RMSE), Mean Squared Error (MSE), R2 Metric (R2) and Mean Absolute Error (MAE). We also denote the execution time in seconds. Averaged over 3 repetitions where each run of Sp-RST has 10 iterations.

7 Conclusion

Making use of powerful data mining techniques, distributed infrastructures and massive data sets, which are provided by international organizations, is of primary importance to assist epidemiologists in their analytical studies of public interest. In this paper, we have presented a case study for using a Rough Set theory approach as a data mining technique in the context of epidemiology. Our study uses a previously introduced distributed method called Sp-RST and considers a data set provided by the Open Cancer organization. After some data preprocessing we perform feature (risk factor) selection with Sp-RST and analyze the results from two different angles: insights epidemiologist can gain from the selected risk factors and the quality of the regression model. From our analyses, we conclude that using feature selection in the considered context is highly beneficial. The data set obtained is much easier to interpret and still yields comparable regression results. The process of regression is much faster on the reduced data set.

As discussed earlier, we are currently only considering a subset of the OpenCancer data set. We plan to expand our study to the complete set in future work. Moreover, we will work more closely with epidemiologist to further improve our method, both with respect to interpretation of the results and precision of the regression model. One important aspect in this context will be the consideration of missing values in the original data set.

References

1. Pawlak, Zdzisław and Skowron, Andrzej: Rudiments of rough sets. Information sciences **177**(1), 3–27 (2007)
2. Bagherzadeh-Khiabani, Farideh and Ramezankhani, Azra and Azizi, Fereidoun and Hadaegh, Farzad and Steyerberg, Ewout W and Khalili, Davood: A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. Journal of clinical epidemiology **71**, 76–85(2016)
3. Mooney, Stephen J and Westreich, Daniel J and El-Sayed, Abdulrahman M: Epidemiology in the era of big data. Epidemiology (Cambridge, Mass.) **26**(3), 390(2015)
4. Woodward, Mark: Epidemiology: study design and data analysis. CRC press (12013)

5. Zaineb Chelly Dagdia and Christine Zarges and Gaël Beck and Mustapha Lebbah: A distributed rough set theory based algorithm for an efficient big data pre-processing under the spark framework. In: Proceedings of the 2017 IEEE International Conference on Big Data, pp. 911–916. IEEE, Boston, MA, USA (2017)
6. Thangavel, K and Pethalakshmi, A: Dimensionality reduction based on rough set theory: A review. Applied Soft Computing **9**(1), 1–12 (2009)
7. Amersi, Farin and Agustin, Michelle and Ko, Clifford Y: Colorectal cancer: epidemiology, risk factors, and health services. Clinics in colon and rectal surgery, **18**(3), 133 (2005)
8. Banerjee, Amitav and Chaudhury, Suprakash: Statistics without tears: Populations and samples. Industrial psychiatry journal, **19**(1), 60 (2010)
9. Porta, Miquel: A dictionary of epidemiology. Oxford University Press (2008)
10. Dicker, Richard C and Coronado, Fatima and Koo, Denise and Parrish, R Gibson: Principles of epidemiology in public health practice; an introduction to applied epidemiology and biostatistics. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention (CDC) (2006)
11. Liu, Huan and Motoda, Hiroshi and Setiono, Rudy and Zhao, Zheng: Feature selection: An ever evolving frontier in data mining. Feature Selection in Data Mining, 4–13 (2013)
12. Schneider, Johannes and Vlachos, Michail: Scalable density-based clustering with quality guarantees using random projections. Data Mining and Knowledge Discovery, 1–34 (2017)
13. Dean, Jeffrey and Ghemawat, Sanjay: MapReduce: a flexible data processing tool. Communications of the ACM, **53**(1), 72–77 (2010)
14. Zhai, Tingting and Gao, Yang and Wang, Hao and Cao, Longbing: Classification of high-dimensional evolving data streams via a resource-efficient online ensemble. Data Mining and Knowledge Discovery, 1–24 (2017)
15. Vinh, Nguyen Xuan and Chan, Jeffrey and Romano, Simone and Bailey, James and Leckie, Christopher and Ramamohanarao, Kotagiri and Pei, Jian: Discovering outlying aspects in large datasets. **30**(6), 1520–1555 (2016)
16. Zhang, Jianfei and Wang, Shengrui and Chen, Lifei and Gallinari, Patrick: Multiple Bayesian discriminant functions for high-dimensional massive data classification. Data Mining and Knowledge Discovery, **31**(2), 465–501 (2017)
17. Pawlak, Zdzisław: Rough sets: Theoretical aspects of reasoning about data. Springer Science & Business Media (2012)
18. Shanahan, James G and Dai, Laing: Large scale distributed data science using apache spark. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2323–2324, ACM (2015)
19. Polkowski, Lech and Tsumoto, Shusaku and Lin, Tsau Y: Rough set methods and applications: new developments in knowledge discovery in information systems. 56, Physica (2012)
20. Guller, Mohammed: Big Data Analytics with Spark: A Practitioner's Guide to Using Spark for Large Scale Data Analysis. Springer (2015)